DOCUMENT RESUME

ED 309 773 IR 052 832

AUTHOR Beam, Paul; Huntley, Frank
TITLE "M" to "Moonless": Lexical Databases in Development.
PUB DATE Jul 87
NOTE 13p.; Paper presented at the International Conference on Data Bases in the Humanities and Social Sciences (July 11-13, 1987).
PUB TYPE Viewpoints (120) -- Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computer Assisted Instruction; Computer Software; Computer System Design; *Databases; Dictionaries; English Literature; Foreign Countries; Higher Education; *Information Retrieval; *Lexicography; *Online Systems; Research Tools
IDENTIFIERS *New Oxford English Dictionary; University of Waterloo (Ontario)

ABSTRACT

This paper describes the characteristics of lexicographic software programs used in a module on Alexander Pope's "The Rape of the Lock," a major component of a course on computer-assisted learning (CAL) at the University of Waterloo in Ontario, Canada. Two of the databases are specific to that school and are small, individualized, and frequently adapted in response to student request. The other is the New Oxford English Dictionary (NOED), which is being developed jointly by Oxford University Press and the University of Waterloo. At the time of this study, the only segment of NOED available for use was the alphabetical range of M through Moonless. This became a prototype of one of three sources of research for students using the CAL module, the other two being a bibliography and a lexicon specific to an understanding of terms in the context of Pope's work. The NOED has a projected 500,000 main headwords, 2.2 million citations, and 425,000 cross-references, and cannot be altered or annotated. Because accessing and managing the information from an unrestricted dictionary of this size can be intimidating for the naive user (i.e., the student), the two smaller database systems were assimilated into files extracted from NOED which are easily modifiable by instructors. Also discussed are questions about defining user need, the sophistication level of users, and awareness, and the overall effectiveness of lexical materials. (SD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED309773

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

# "M" to "Mooniess": Lexical Databases In Development

A paper presented for the International Conference on Data Bases in the Humanities and Social Sciences.

July 11-13, 1987.

Paul Beam Frank Huntley

University of Waterloo,Ontario

Running head: "M" to Mooniess

IR053832

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Paul Beam

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."



BEST COPY AVAILABLE

Abstract

In May of 1985 as a major component in a course on computer-assisted learning at the University of Waterloo, we commenced work on a module of Alexander Pope's The Rape Of The Lock. In our initial plans it was apparent, from the selection of this particular work, that students would benefit from as wide a range of lexical aids as we could provide. At the same time a very limited part of the New Oxford English Dictionary became available for research on campus. We could gain access to only the quotations in the alphabetical range of the first part of the letter "M". This became a prototype of one of three sources of research for students using the CAL module, the other two being a bibliography and a lexicon specific to an understanding of terms in the context of Pope's work. These latter are easily modifiable by instructors, while the NOED material, extensive and authoritative, is exclusive to the University and Oxford University Press, and admits of no student or teacher modification.

Since all three databases are available on a series of function keys when the user places the cursor on a word or phrase, they are convenient beyond any conventional textual expression. We wish to consider the major differences for student learning and scholarship between the NOED, on the one hand, and the two smaller, quite individualized, databases on the other.

"M" to "Moonless": Lexical Databases In Development

To the lexicographer and the language specialist questions of comprehensiveness are pre-eminent. Consequently, speed of response and appropriateness of the material to an audience of students or naive users are secondary concerns because, in the overall of research, these users of a database are poor cousins whose problems can be presumed to be addressed after the greater concerns of total information have been met. This may be an oversimplification but it points to the problems addressed in this paper.

In 1857, at the inception of the New English Dictionary, the forerunner of the Oxford English Dictionary, no consideration was given to potential problems of information overload for unfortunate university undergraduates. In a world of paper all information has to be present and viewable in one place and the luxury of discretionary or selective definitions does not apply; things have to be there or they cannot be known. Printing, binding and distribution costs prohibit layered or discretionary editing of what a given reader is shown. This concept has become possible only with the kinds of software becoming available in projects like the New Oxford English Dictionary (NOED), a joint undertaking by the Oxford University Press and the University of Waterloo (1).

While the abilities to search an item or set of correlatives by date, author, location or grammatical structure are hailed as unparalleled benefits by scholars, the idea of limiting what other users might access may be an alarming prospect to some; it smacks of censorship, no matter how beneficently it is intended or how well it may be designed. Like the search features of this new software, it is a recent addition to lexicography and it can now begin to be investigated for the first time. It will remain a

retrograde tool at best, but we believe its value will become increasingly clear as teachers especially attempt to have their students benefit from a database as comprehensive and elegant as the NOED in its final form.

To give some sense of scale, the projected number of entries is about 500,000 main headwords with lemmas, 2.2 million citations and 425,000 cross-references. This material is being augmented even as it is being prepared. In person-years, as another index, one hundred and twenty typists have worked eighteen months to enter the main corpus of materials which comprises 500 million keystrokes and 350 million characters.

The system's size increases further to fill software needs. The NOED requires nearly as much storage for the annotated tags to search the text as for the components of the definitions themselves. While very efficient text-processing algorithms exist to permit rapid location of information, they are only implemented on the NOED's local development machine, a SUN workstation with a UNIX operating system, and are unavailable to us. However, the University's CMS/VM system, available to all students, provides a file organization which allows us to rebuild quickly a hierarchical database and thus quickly access the correct file, while our chosen language, REXX, with its powerful string-handling capabilities and explicit editor control (including response trapping), speeds search and extraction time for the appropriate records. NOED material cannot be altered or annotated by us, so our only concern has been to provide a simple, convenient connection to it. Behind that provision lie the unresolved questions about which users benefit, and how, and what degree of exposure up to the entire definition of a particular word, or the selection of parts of the definition that might be extracted and displayed.

The two internal dictionaries give no problems in access time because the are small and specific. They are created, modified and expanded by us in response to student requests, as distinct from the NOED's comprehensive coverage of a language and all of its users. Students using the lexicon specific to a lesson now have easy entry to permit the creation of a great deal of this material themselves, and so our search efforts are directed to the output from the database - how to search most effectively and extract the needed information. The real power of the software tools for the NOED requires an experienced and knowing hand to be used effectively. The problem in an application to an educational situation is a kind of obverse of this wealth of information; this might be illustrated best by an example from a text students care to study and analyse critically. One objective is to make them intensely aware of words and their multiple meanings in different contexts. In a poem entitled "Afterwards", Thomas Hardy commences a bleak speculation with the line, "When the Present has latched its postern behind my tremulous stay". One would think that vocabulary concerns might center on terms like "postern" and "tremulous", but these words are resolved easily when the distinction is made that the one is a back-gate giving egress from an enclosed area onto open or uncultivated land, usually from the rear of a structure like a court or farm yard, while the other need only be distinguished from the more-to-be anticipated "trembling". The real sleeper here is the simple word, "stay". When a student attempts to access its definition from the NOED by placing the cursor under it as the passage is displayed, the result, in its un-edited form, is some ninety-six screens of information. These results are not what was ever intended, either as scholarship by the researcher or for learning processes in which the instructor wants to assist. This is what some users have come to fear from

such a dictionary with its ability to store and present huge quantities of data. From a researcher's perspective, such problems need not arise in an electronic format because, like the organizing software of the database, the scholar's skill is in getting directly at the essence of the material sought by eliminating from the search all extraneous data. "Stay" in the context can be only a noun as its adjectives and location relative to the preposition "behind" verify. If the passage's date were known, more modern meanings could be ignored (though we are aware of over-simplifying for our argument). By processes of elimination like these, the experienced reader may reduce an overwhelming structure to a manageable and revealing display and may draw important conclusions from it.

What hope this for our poor students who use this tool presumably because they lack precisely these skills and are now attempting to learn to acquire them? An unrestricted dictionary of this size is formidable and potentially misleading. Among nominals of "stay" are ropes and cables used to secure masts and rigging of sailing vessels, so grammar restrictions alone will not protect the unwary against that misreading. And, arguably, in many cases it should not. Obviously, in any language training, students must be encouraged to experience, in Blake's admonition, "Enough, or too much". Still, ninety-six screens for a four-letter word must inevitably daunt the most dogged and frustrate the most clear-sighted without much advancement of learning for either.

Our solution is a compromise for the moment. At some future time it may be possible, working from an electronic database, to design a screening device set by the instructor, or even the learner, appropriate to the user's maturity, sophistication and research needs. Such a device must have in it

some of the elements of a real-language parser, as well as the software described above. For the present, we offer this solution, one which has worked very well for The Rape of the Lock. Among a range of function keys one is assigned to provides the NOED definition of the word in its complete expression. Its mate provides the instructor's definition in a small lexicon exclusive to the module itself.

The virtues of this second dictionary are three. The instructor defines, or has defined, a meaning appropriate to the subject, within the context of the work. This definition has the advantages of brevity and direct application to the context of the passage. Secondly, the material is easily entered and modified in light of users' responses so that it becomes increasingly relevant to the subject. Finally, at any point the instructor may turn over editorial authority to students to modify or expand parts of the database.

## Implementation

Our problems required different algorithms and technologies than those provided by the traditional textbases or databases. We were required to coordinate a textbase over which we had no control with a second textbase that was exclusively under our control. Access for hundreds of concurrent users was necessary, which limited our choice of computer hardware to a large mainframe. The users, who were both querying the data and adding to the database, were naive, precluding the use of a traditional database query language. Finally, the project was a small component of a larger textbase project, and so development was limited. Thus, we had to provide a simple, fast and familiar system on a large mainframe.

By implementing the system on our mainframe running under CMS/VM,

and by selecting an appropriate language, REXX, a powerful editor, XEDIT, and some system commands, we provided a quick, consistent interface to the lexicons which permits easy modification and addition of new materials as required. The following four points illustrate the use of these tools in implementing both lexicons.

The CMS system has a "listfile" command that can quickly sort filenames. Since the NOED is set up as an alphabetically ordered group of files with significant titles, it can be loosely conceptualized as an reconstructable, hierarchical database. A "hash table" can be built with the "listfile" command, which quickly sorts the NOED filenames into a file. This provides a table of alphabetical "keys" and allows the search range to be narrowed quickly with the selection of a single file, such as the file containing the definitions from "M" to "Mooniess".

The system's editor, XEDIT, is very fast and can be controlled from within the programming language, REXX. Once the file storing the word is selected, XEDIT locates the word and then displays the definition. At the same time, it gives us fast and efficient control of the main text, because XEDIT is the central editor on a large mainframe; its capabilities have been optimized. Since we control XEDIT, we have fast, full-screen control of all text, databases and user responses. As well, since our system uses the editor for controlling text, students can use it in their work on the smaller lexicons, and are thus provided with a familiar interface with which they may add their research work to The Rape Of The Lock database.

Thirdly, REXX's powerful string manipulation commands allow manipulation of both database information for screen processing and user responses for system control. The string manipulation is necessary for

handling:

1) user-responses,
2) the text of The Rape Of The Lock,
3) internal lexicons,
4) the NOED information itself.

For example, the following statement removes all non-numeric characters from the string variable 'answer':

```
answer = translate (answer,'',',./?"";:!{}\^@#$%*()-_=+|`~<>')
```

Finally, we ensured the possibility of future developpment by choosing REXX as the language. Rexx is a structured language like Pascal, with high-level language constructs such as the "Do while" statement, which permits easy development and modification by experienced programmers.

The conveniences of size and speed inherent in a main-frame system are only slightly qualified by problems of access. CMS can handle over five hundred simultaneous users and many of our students have modems so they may access the module and dictionary facilities from any dial-in point. They quickly come to take the new environment for granted. A number have impressive cybernetics skills and are valuable additions to our research. :fn id= ref1. Across the range of CAL projects are inter-active writing aids accessed by students in all faculties on the main CMS system, a lesson-creation module for teachers in QNX for the Ministry of Education of Ontario and a language training module in Software A.G.'s Natural. :efn.

## The Future

What we most lack are two features to assimilate the discreet systems just described. It is improbable that the owners of the NOED will ever permit any or all users editorial access to the data of the dictionary; nor do we feel they should. So, for our pedagogic needs, some process by which

limited excerpts could be brought into discreet CAL modules and then be edited "on-site" by instructors for a particular lesson, would be very advantageous. The need for remote access would be reduced, appropriateness of subject matter for users would be retained and instructor efficiency would remain high. This process would blend the scope of the large dictionary with the instructor's editorial capacity to present material appropriate to the content and the user. The solution to this problem is not so much a technical as a legal one, requiring resolution of potential copyright infringements.

The second concern is larger, but of more academic interest. We are considering it only as the large electronic lexicon becomes complete and available. Questions about how to define a particular user, how to assess levels of sophistication, awareness and need, are beginning to emerge among the responses we receive from the student work on The Rape Of The Lock. The information users seek, how effective they find it and how they use it assists in defining the kind of editorial screen appropriate to this group. We measure need by considering which words are accessed. Effectiveness of the lexical material is judged in part by user comments sent in response to definitions; some of these communications are complex and simulating. How students use the definitions is a difficult variable to define; it appears more as a shadow than fact. The ways in which students apply lexical and semantic insights in their written expressions seem to be the best index but this is at two removes from the material itself. In all circumstances in which we would attempt to restrict the quantity of lexical material for display the entire definition would be available, should the student request it. But evidence is emerging to permit development of the definition screen. This forms the basis for our proceeding into what we regard as the merging

for benefits of two distinct lexical philosophies.

(1) Information and statistics about the NOED and work at the University of Waterloo have been provided through the courtesy and research of Professor Delbert Russell of the Department of French and a Director of the project at this university. Our thanks for his consideration and assistance.